Investigating the effect of volume, rating, and valence of product reviews on helpfulness

Auteur : Carnevali, Gauthier
Promoteur(s) : Ittoo, Ashwin
Faculté : HEC-Ecole de gestion de l'ULg
Diplôme : Master en sciences de gestion
Année académique : 2015-2016
URI/URL : http://hdl.handle.net/2268.2/1892

HEC ULg
Ecole de Gestion de l'Université de Liège

Investigating the effect of volume, rating, and valence of product reviews on helpfulness

Jury :
Promoter :
Ashwin ITTOO
Reader(s) :
Alessandro BERETTA
Michael GHILISSEN

Dissertation by
Gauthier CARNEVALI
For a Master's degree in Management

Academic year 2015 / 2016

Table of contents

1 Introduction 4

2 Literature 8

2.1 MIS Quartely 8

2.2 Electronic commerce and research applications 9

2.3 Automatically assessing review helpfulness 11

2.4 Importance of online product reviews from a consumer's perspective 12

3 Methodology 14

3.1 Data 14

3.2 Samples 15

3.3 Selecting variables 17

4 Descriptive statistics 22

4.1 Polarity and Flesch reading ease 22

4.2 Rating and text difficulty 26

5 Hypothesis 28

5.1 Experience goods 28

5.2 Search goods 28

6 Statistical method used 30

7 Assumptions 32

7.1 Assumption #1 32

7.2 Assumption #2 32

7.3 Assumption #3 32

7.4 Assumption #4 33

7.5 Assumption #5 33

7.6 Assumption #6 36

7.7 Assumption #7 37

7.8 Origin of outliers 38

7.9 Decision concerning assumptions 41

8 The models 42

8.1 Experience good: Books 42

8.2 Summary of experience good models 45
8.3 Search goods: Cell Phones 46
8.4 Summary of search good models 48
8.5 Hypothesis results 49

9 Conclusion 50

10 Annexes 54
10.1 Syllable code 54
10.2 Descriptive statistics 55
10.3 Reviews 56
10.4 Outlier table 58
10.5 Models 63

11 References 66

12 Bibliography 70

1 Introduction

Over the past decade the number of internet reviews on products has exploded. Amazon, in particular, has had a major increase in reviews over the past years. This has led to consumers having a surplus of information and finding it difficult to find reliable information on products.

.

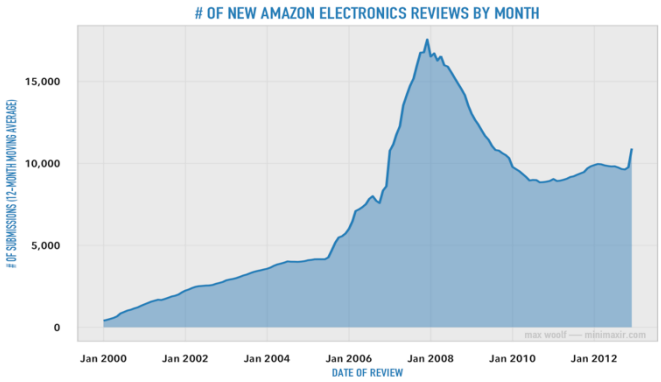

Figure 1: shows increase in electronic reviews

Max Woolf (2014)

Consumers find themselves making decisions based on incomplete and sometimes false information.

> *"Consumer know that seeking this information is costly and time consuming, and that there are trade-offs between the perceived costs and benefits of additional information (Stigler 1961)."*
>
> *"Therefore, the total cost of product must include both the product cost and the cost of search (Nelson 1970)"*

Making sure that consumers find the information their looking for without searching or at least minimal searching will **improve consumer experience** and may make the difference between a company and its competition. As a matter of fact, Kumar and Benbasat (2006) found that **customer reviews heavily influence online product choices**.

According to Georg Lackermair, Daniel Kailer, & Kenan Kanmaz (2013). What consumers are looking for are reviews which give both the pros and the cons of a product. They wish to compare positive and negative reviews with each other.

Using reviews to choose what product to buy has become second nature nowadays. Whether the purchase is done online or offline, reviews are usually a **primary source for**

product information. A recent study, Dan Hinckley (2015), found that **67,7%** of potential customers are influenced in their purchase by online reviews. 29,6% said that reviews are very or absolutely important in making their purchase decision.

For these reasons it has become important for online companies like Amazon to be able to identify helpful reviews from unhelpful reviews. Putting helpful reviews at the forefront will boost customer satisfaction.

Another, and probably, one of the most important reasons why reviews are important is that they can contribute to an increase in sales. Here are a few interesting statistics on reviews by Graham Charlton (2015):

- 50 or more reviews per product can increase conversion rates by **4,6%**.
- **63%** of customers are more likely to buy from a website where there are customer reviews.
- Reviews written by consumers are significantly more trusted than information that comes from a manufacturer.
- Reviews produce on average **18%** uplift in sales.

The goal of this thesis is to identify some of the variables that make a review helpful or unhelpful. The title of the thesis states that volume, rating and valence will be used as the independent variables.

Volume in this case corresponds to text difficulty. The text difficulty formula takes into account number of words, number of sentences, and number of syllables. Rating corresponds to the number of stars a reviewer gives a product. Finally, valence is the polarity of the text.

The different samples will be divided into two major categories: **experience and search goods**. This distinction is done in order to see if review helpfulness is influenced differently according to the product type.

Furthermore, hypothesis will be stated on how each independent variable will react according to the product category for example, we emit the hypothesis that experience goods are more likely to have helpful reviews if star ratings are low.

The thesis is structured as stated below.

The first chapter focuses on available literature. We will present the methodology and results of different papers. The literature chosen are texts that influenced our own methodology.

The second part focuses on the methodology used. Sample segmentation and the identification of explanatory variables are the main focus of this chapter. Explanation on how polarity and text difficulty were extracted from the data is given in this part.

Chapter four focuses on descriptive statistics. You can find some general information on the different variables. We will focus also on one of the main problems for the Flesch reading ease test with online reviews.

Chapter five and six are two short chapters. The first one's main purpose is to define the different hypothesis concerning the variables. Chapter 6 introduces the statistical method used to create the different models. The statistical method that was used was a binary logistic regression.

Chapter seven tests the assumptions required in order to do a logistic regression. This chapter also focuses on outliers and their origin.

In chapter eight we run the logistic regression and give the results. Furthermore, hypothesis are tested at the end of this chapter. The program used to run the regression was IBM's SPSS statistic.

The final part is the conclusion. We summarize results in this chapter, give recommendations on how to improve the model and possible practical applications of the model.

2 Literature

In this chapter I will present different research papers that had a similar research question and who's influence was greatest in writing this paper. A lot of the literature on this particular subject is very similar from one paper to another we tried to choose texts that had different approaches and which took into account different variables.

2.1 MIS Quartely

This paper written by Susan M. Mudambi and David Schuff (2010) posed the following question: "What makes a helpful online review? A study of customer reviews on Amazon.com?". They, like me, make the distinction between experience and search goods. They used MP3 players, music CD, and PC video games as their experience goods and cell phones, digital cameras, and laser printers as their search goods. They had a total data set of 1587 products.

Before creating their model, they posed the three following hypothesis:

1. ***Product type*** *moderates the effect of review extremity on the helpfulness of the review. For experience, goods reviews with extreme ratings are less helpful than reviews with moderate ratings.*
2. ***Review depth*** *has a positive effect on the helpfulness of the review*
3. ***The product type moderates the effect of review depth*** *on the helpfulness of the review. Review depth has a greater positive effect on the helpfulness of the review for search goods than for experience goods.*

They then went on to define their different variables. The helpfulness is of course the dependent variable. They used review extremity, review depth, and product type as their explanatory variables.

Review extremity was measured using star ratings. Review depth is represented through number of words and product category is a binary variable where products are either search or experience goods. They also added a control variable, the total number of votes. A review with a 100 votes saying it is helpful may not have the same interpretation as a review with one vote.

They used a Tobit regression to analyze their model and used the likelihood ratio and Efron's pseudo R-square value to measure goodness of fit.

Table 4. Regression Output for Full Sample				
Variable	Coefficient	Standard Error	t-value	Sig.
(Constant)	61.941	9.79	5.305	0.000
Rating	–6.704	7.166	–0.935	0.350
Rating²	2.126	1.118	1.901	0.057
Word Count	**0.067**	**0.010**	**6.936**	**0.000**
Product Type	**–45.626**	**12.506**	**–3.648**	**0.000**
Total Votes	–0.0375	0.022	–1.697	0.090
Rating × Product Type	**32.174**	**9.021**	**3.567**	**0.000**
Rating² × Product Type	**–5.057**	**1.400**	**–3.613**	**0.000**
Word Count × Product Type	**–0.024**	**0.011**	**–2.120**	**0.034**

Figure 2.1: Results of regression

Their results supported all three of their hypothesis.

2.2 Electronic commerce and research applications

Written by Nikolaos Korfiatis, Elena Garcia-Baricocanal, & Salvador Sanchez-Alonso. (2012), this research paper is about "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content.". This article tries to evaluate the relationship between helpfulness, rating score and the qualitative characteristics of review text.

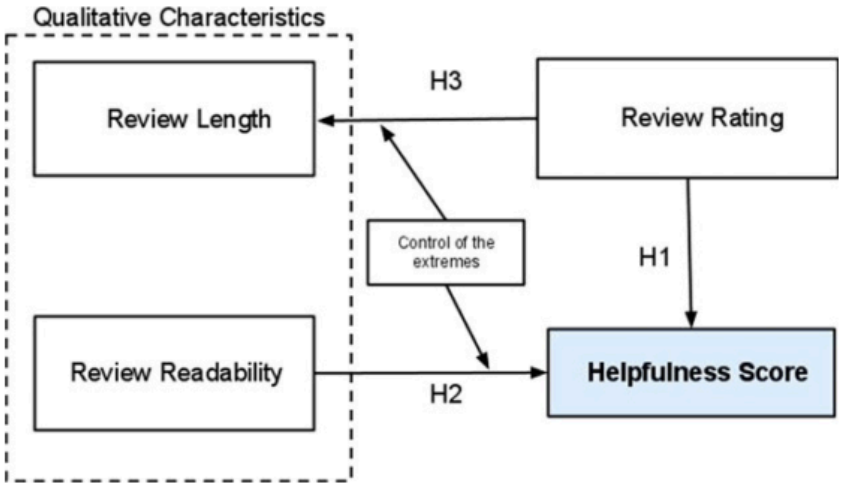

Figure 2.2: representation of their model

Figure 2.2 depicts the model they used in their study. Three main hypothesis were made:

1. **a.** *The helpfulness of a review is directly affected by its rating*
   b. *The helpfulness of a review is* ***unaffected*** *by its rating*
2. **a.** *The helpfulness of a review is directly affected by its qualitative characteristics, and in particular, by the readability of the review text.*
   b. T*he* ***extreme*** *helpfulness of a review is directly affected by its qualitative characteristics, and in particular, by the readability of the review text.*
3. **a.** *The length of the review text is affected by the value of the rating given by the review.*
   b. *The length of the review text is affected by the* ***extreme*** *value of the rating given by the review.*

As stated earlier the models different variables are: rating, number of words, and qualitative test evaluations.

In order to evaluate the quality of text they used four different readability tests. Using this many different tests helps parrying shortcomings that could be found if only using one individual test.

They, like in the previous research paper, used a Tobit regression to evaluate their model. The model gave significant results and, apart for one, supported their hypothesis.
The hypothesis that didn't fit was hypothesis 1b "*The helpfulness of a review is* ***unaffected*** *by its rating"*.

From their data analysis they found support for the following statements:

- High helpfulness of a review is affected by its positive rating value.
- Highly helpful and extremely helpful reviews contain more readable text than reviews that are less helpful or not helpful.
- Reviews are longer when they are positive or absolutely positive and shorter when they are negative and absolutely negative.

2.3 Automatically assessing review helpfulness

This paper was written by Soo-Min Kim, Patrick Pantel, Tim Chklovski, & Marco Pennacchiotti. (2006). They try to create a model capable of automatically classifying a review as helpful or unhelpful based on certain criteria.

This article significantly deviates from the two previous papers through their choice of explanatory variables. It is one of the papers with the most explanatory variables. They defined their independent variables into five broad categories.

Table 2.1

Category	Variable	Explanation
Structural features	Length	Total number of words
	Sentential	Observations of the sentences: number of sentences, average length, question sentences, exclamation marks.
	HTML	Bold tags
Lexical features	Unigram	*tf-idf* (reflects the importance of the word in the review) statistic of each word occurrence.
	Bigram	*tf-idf* statistic of each bigram (two words) occurring in the review.
Syntactic features	Syntax	This relates to the percentage of tokens that are nouns, verbs, adjectives, and adverbs.
Semantic features	Product features	These are words that relate to a product. Ex. "Capacity" for MP3 players or "zoom" for cameras.
	General inquirer	Number of positive and negative words.
Meta-Data features	Stars	Overall rating of product, star score given by reviewer.

For their regression model they used the SVM regression tool.

Their results showed that unigram outperformed bigrams. Bigrams give as good a result because reviews are generally short. The unigram also subsumes the semantic features.

The star rating is also significant whether the simple star rating is used or the difference between star rating and the average rating of review is used doesn't seem to change the result significantly.

Structural features, sentential, HTML, and syntax features did not seem to have significant influence. These features didn't improve their systems performance.

2.4 Importance of online product reviews from a consumer's perspective

This research paper was written by Georg Lackermair, Daniel Kailer, & Kenan Kanmaz (2013). As the title indicates the researchers try to take the perspective of consumer's rather than simply using explanatory variables that past research used.

They make the following hypothesis:

1. *Online reviews **increase trustworthiness***
2. *Reviews are an important source of information for online transactions.*
3. ***Users are willing to contribute:** user generated reviews and ratings can be classified as services; whose quality improves as the contributions of users rise.*

In order to check for their hypothesis, they used a survey. They found that user reviews are in fact important for the decision making process of a consumer. Consumers search for reviews where there are both positive and negative aspects of a product.

They argue that the decision making process for online purchasing is divided into phases. In the early phase the consumer tries to narrow down products to try and find products that match his/her requirements, online reviews are important for this phase. They further argue that consumers need compact and concise information related to products. Reading a lot of reviews isn't very efficient. Consumers want quality information rather than quantity.

3 Methodology

3.1 Data

The data we used comes from Amazon. We collaborated with a researcher named Julian McAuley (2015) who works at the University of California, San Diego. He had collected all the reviews on Amazon for products spanning from 1996 to 2014.

Duplicates in the data were removed by Julian McAuley. Users with more than one review for a same product or who have more than one account who reviewed a same product with both accounts were removed. The duplicates represent less than one percent of total reviews.
The following information was collected for all product reviews:

- Reviewer ID
- Product ID
- Reviewer name
- Helpfulness rating of the review
- Review text
- Overall rating of the product
- Summary of the review
- Review time

We used the following elements for the analysis: helpfulness, review text, and overall rating of the product. We extrapolated other variables from the data such as text difficulty and polarity of the review.

3.2 Samples

3.2.1 Search and experience goods

One of the challenges was to create representative samples out of all the collected data.
We decided to create four major samples with two different large categories of products. The first sample is composed of search goods while the second is composed of experience goods.

Quite a few research papers used this kind of segmentation (Susan & David, 2010; Nikolaos et al., 2012; Philip (2011). They usually found that explanatory variables acted differently according to which category the product is in.

This categorization may not be enough, Amazon being able to commercialise the long tail, they have a huge number of niche products. A lot of products nowadays are hard to define in one category. Take for example music. At first glance it should be categorized as an experience good, you have to listen to it to form an opinion. However, through reading critics and marketing an opinion may be formed beforehand.

3.2.1.1 Experience goods

Experience goods are products where quality is hard to evaluate because it strongly depends on a person's taste. A good example of an experience good is food. One must buy it and consume it in order to evaluate quality.

Nelson (1970) defined an experience good as being a product where the consumer, in order to evaluate the quality and utility, had to buy and consume the product. The consumer can't base his purchase on pre-existing information on product quality.

Two different experience goods were used; books and video games. These products usually have to be bought and used before being able to form an opinion.

3.2.1.2 Search goods

Search goods are products were it is possible to obtain information on product quality. They have key characteristics that allow you to measure, to a certain degree, the quality of a product. An example of a search good could be a USB drive, we can identify the following attributes: storage space, speed of data transfer, and physical size.

> *"…relatively easy to obtain information prior to interaction with the product: key attributes are objective and easy to compare, and there is no strong need to use one's senses to evaluate quality. Susan M. Mudambi (2010)."*

The two samples that we used for this category were electronics and cell phones. These products usually have well defined criteria and quality can be measured before purchase.
The electronics sample is primarily composed of headphones, MP3 players, cameras, etc.

3.2.1.3 Thinning the data

The size of the data was extremely big (60g of data in total), because of this we had to reduce the quantity of reviews so that the computer could process the data.

We didn't take into account data with 5 or less votes for helpfulness. Not enough people have voted these reviews to make the helpfulness rating significant.
The data we used was already divided in broad categories, we took from each category 30000 samples at random. Details of the sample can be seen in the table below.

Table 2.1

	Starting sample	Reviews with 0-5 votes	Final N
Books	30000	25285	4715
Video games	30000	26045	3955
Cell phones	30000	25491	4509
Electronics	30000	25136	4864

3.3 Selecting variables

3.3.1 Dependent variable

The outcome variable is helpfulness. The goal of this thesis is to identify what influences the helpfulness rating and whether it's possible to predict the outcome. In order to make the distinction between helpful and unhelpful reviews we defined a threshold at which point a review becomes helpful or unhelpful.

Although Scott Bolter (2013) defined the threshold at 70% we will define ours as being at 60% based on a recommendation from Ashwin Ittoo (personal communication, 11March 2016).

3.3.2 Independent variable

There are numerous explanatory variables that can be used to explain a products helpfulness, however we will be using only 3 variables.

3.3.2.1 Polarity of text

The polarity of a review measures if the text is negative or positive in retrospect to the product. In order to measure this variable, we counted the number of positive and negative words and then gave the review a score by subtracting the negative words to the positive. So if a review had, for example, 5 positive words and 10 negative words its polarity score would be -5.

The words that were used for the search in the text come from Bing Liu's opinion lexicon. The lexicon contains 6789 words in total. We chose this lexicon because it has an important number of words and it is easy to use.
The lexicon does have limits to it. There are more negative words compared to positive words, which could increase the chance of a text being negative rather than positive.

3.3.2.2 Text difficulty

Michael P.O'Mahony and Barry Smyth (2010), wrote a small paper on how the readability of a review influences helpfulness. They found that different readability tests gave more or less the same results for certain goods while for others it didn't. They hypothesize that if a text is too hard or too easy to read then the review is less likely to be helpful.

Other research, Susan M. Mudambi & David Schuff (2010), showed that review depth (length of review) explained review helpfulness. Nikolaos Korfiatis, Elena Garcia-Baricocanal, & Salvador Sanchez-Alonso (2012), used both length of review and readability as explanatory variables. They found that readability (text difficulty) had a bigger influence on helpfulness than review length.

We will only use text difficulty. Because the Flesch-Kincaid test uses number of words there is a chance of correlation between text difficulty and number of words if we used both variables.

The text difficulty is measured using a **Flesch-Kincaid test,** this test gives the text a grade between 1 to 100, with one being hard and one hundred easy. Although sometimes texts can go beyond 100 or under 1. The following formula is used to give the score:

$$f(x) = 206.835 - 1.015\times\left(\frac{Total\ words}{Total\ sentences}\right) - 84.6\times\left(\frac{Total\ syllables}{Total\ words}\right)$$

The test is based on number of words and number of syllables. The more words and the more syllables the harder the text. Obviously this test is not without flaws. It is far from perfect seeing as it only takes into account 2 characteristics, nevertheless it usually gives a good idea of whether a text is difficult or not.

In order to calculate the number of syllables in a text, we had to first understand exactly what makes a syllable. Denis Eide (2014) defines syllables in a word as being equal to the number of vowels.

> *"Every syllable must have a vowel, and every vowel makes a syllable. This means that the number of vowels in a word is equal to the number of syllables."*

However, when counting, one must also be careful not to include silent vowels, for example when two vowels are together or when a silent "e" is placed at the end of a word. The website Phonics on the web gives the following rule in order to count the number of syllables:

1. Count the number of syllables
2. Subtract any silent vowels
3. Subtract one vowel from every diphthong
4. The number of vowels left is the number of syllables

We used a JavaScript code found on Stackeoverflow (2011), the code can be found in annexes along with an explanation on how it works. The code gave accurate results.

A variation of this test exists which gives the result based on a US grade level. The US grade test emphasises sentence length rather than number of word. We didn't choose it because in internet reviews people often do not respect punctuation codes. This results in very long sentences. Furthermore, reviews are relatively short and usually are composed of 7 to 9 sentences, emphasising words over sentences seems to be the better choice.

Once the Flesch Kincaid test was done we then categorized the result on a scale from 1 to 8 with 8 being easy and 1 very difficult.

Table 2.2

Score	Scale	School level	Notes
90-100	7	5th grade	Very easy to read.
80-90	6	6th grade	Easy to read
70-80	5	7th grade	Fairly easy to read
60-70	4	8th and 9th grade	Plain English
50-60	3	10th and 12th grade	Fairly difficult to read
30-50	2	College	Difficult to read
30>	1	College graduate	Very difficult to read

3.3.2.3 Review rating

Out of all variables, this is probably the most popular one. Almost all research papers used this variable. Susan M. Mudambi & David Schuff (2010), hypothesised that:

> *"Product type moderates the effect of review extremity on the helpfulness of review. For experience goods, reviews with extreme ratings are less helpful than reviews with moderate ratings."*

Their results supported this hypothesis.

Other research, Monic Sun (2011), found that products with a high average rating indicates a high product quality, while a high variance of ratings corresponds to a niche product.

The number of stars give you a summary of whether the reviewer found the product good or not.


Super Smartphone!
By Eric M on 28 June 2016
Farbe: Schwarz | Größe: 16 GB | Verified Purchase
Kaum ein Hersteller schafft es Smartphones mit Hardware aus der Versionen.
Doch Sony schafft es es hier ein Smartphone mit guter Ausstattung wenn es mittlerweile schon viele neueren Zx Compact Versionen gi mit sich rum schleppen möchten.

Figure 2.1

Stars	Meaning
1	I hate it
2	I don't like it
3	It's OK
4	I like it
5	I love it

The stars help consumers identify reviews which are negative or positive. If a person is looking for reviews with different point of views all he has to do is look at the number of stars and he can immediately identify a negative review from a positive review.

4 Descriptive statistics[1]

4.1 Polarity and Flesch reading ease

	Books		**Video games**	
	Polarity	**Flesch reading ease**	**Polarity**	**Flesch reading ease**
N	4715.00	4715.00	3955.00	3955.00
Mean	-4.13	59.68	-2.09	64.16
Std. Error of Mean	0.13	0.23	0.12	0.22
Median	-2.00	59.83	-1.00	64.30
Std. Deviation	8.80	15.93	7.27	14.30
Variance	77.40	253.80	52.87	204.73
Skewness	-1.93	-0.16	-1.93	-0.49
Kurtosis	6.20	0.30	8.74	5.243
Range	82.00	135.97	92.00	247.35
Minimum	-62.00	-20.51	-72.00	-72.61
Maximum	20.00	115.47	20.00	174.73
	Cell phones		**Electronics**	
	Polarity	**Flesch reading ease**	**Polarity**	**Flesch reading ease**
N	4509.00	4509.00	4864.00	4864.00
Mean	-1.37	64.23	-0.27	63.46
Std. Error of Mean	0.08	0.21	0.08	0.21
Median	-1.00	64.83	0.00	63.77
Std. Deviation	5.61	14.12	5.35	14.37
Variance	31.49	199.45	28.60	206.59
Skewness	-0.74	-1.54	-0.42	-0.71
Kurtosis	2.39	16.29	2.88	3.50
Range	57.00	273.42	69.00	173.70
Minimum	-34.00	-159.29	-29.00	-49.89
Maximum	23.00	114.13	40.00	123.81

[1] More descriptive statistics in annexes.

4.1.1 Experience goods

4.1.1.1 Polarity

The median is smaller than the mean by half. The median is a better reference as it is less influenced by outliers. The median results are negative for both samples (-2 and -1) but aren't too far from a score of 0. The minimum for both samples are extreme (-62 and -72) the maximum is less extreme (20 for both reviews).

Skewness is good for both samples while kurtosis isn't.

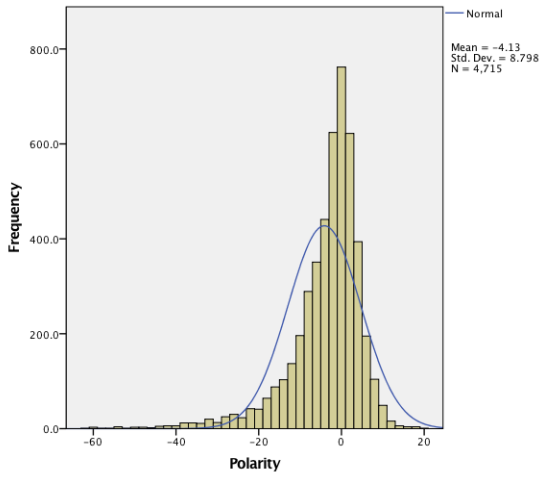

Figure 4.1: Books sample

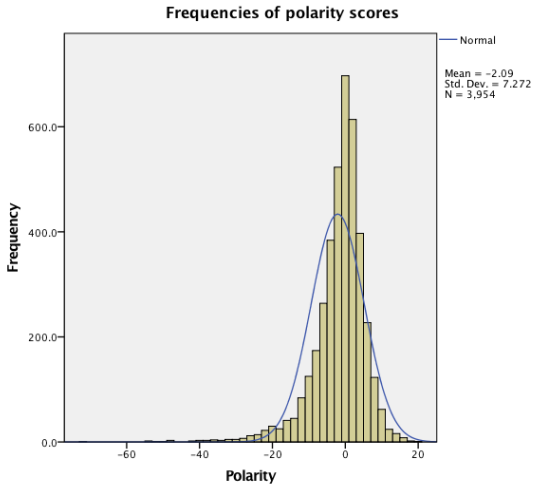

Figure 4.2: Video games sample

Further analysis has shown that extreme negative values over 20 are rare. They represent 3% (video games) to 5% (books) of the data.

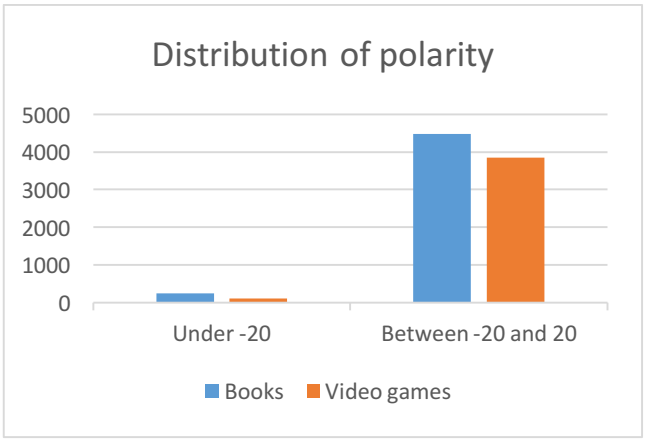

Figure 4.3

4.1.1.2 Flesch reading ease test

NB: We added the Flesch reading test before categorisation to illustrate its limits.

The video game sample has the highest maximum (174.3) and lowest minimum (-72.61). We can further analyse why the Flesch reading ease test gave such a low result and evaluated the text as very difficult (see review 1 in annexes for text).

As we can see the text has syntax and punctuation errors. The high score is due to these punctuation errors. The text is composed of 1 line 309 words and 530 syllables. The high score is due to the number of lines. The first part of the formula divides words by lines while the second part divides syllables by words. If the writer had correctly used punctuation and used "." Instead of ";" then the score would have been more accurate. An increase in number of sentences would have resulted in a decrease in text difficulty. In this case the text is composed of 1 line with 127 words in it.

This demonstrates the main limit of the test. It assumes that the syntax and punctuation of the text is correct and only basis the test on words, sentences, and syllables. While quite accurate for testing books where punctuation and syntax is correct it may sometimes fall short on the internet where anyone can write a review with no regard to punctuation.

There are other examples of reviews in the annex with their test scores.

4.1.2 Search goods

4.1.2.1 Polarity

Texts seem to be mostly neutral with mean and median close to or equal to 0 for both samples.

We can analyse frequencies of values which are bigger or smaller than ± 20. Reviews with polarities over 20 or under -20 are once again marginal. However, unlike with experience goods a few reviews have polarities over 20. These reviews represent 0.8%(cell phones) and 0.4%(electronics) of the samples.

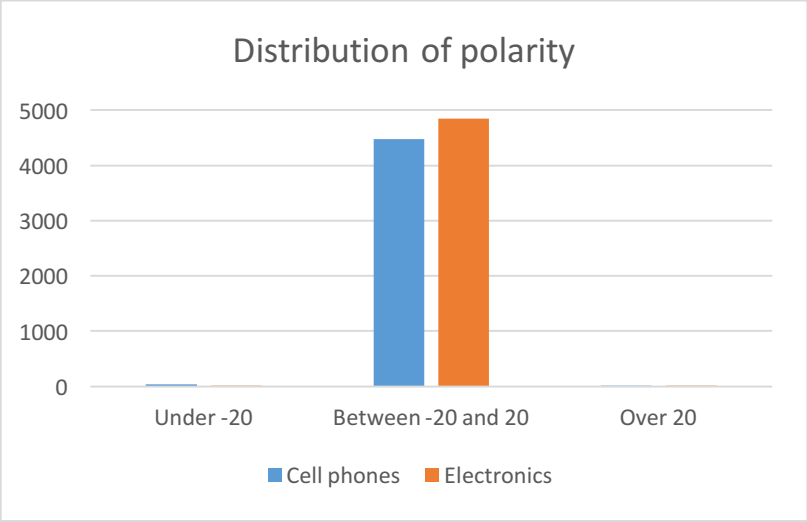

Figure 4.5

4.2 Rating and text difficulty

4.2.1 Experience goods

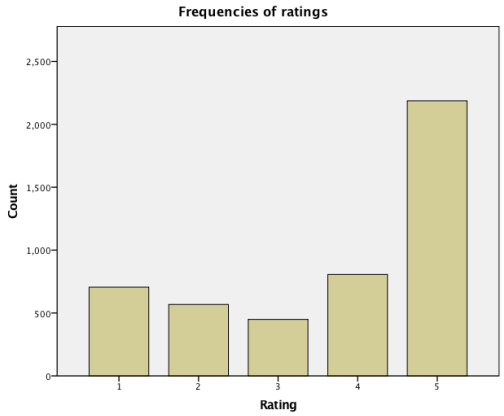

Figure 4.6: Books sample

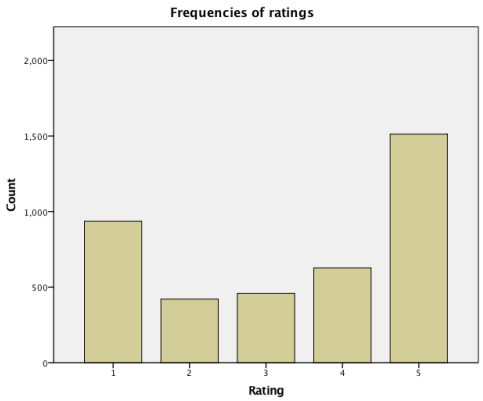

Figure 4.7: Video games sample

Ratings of 5 seem to be most popular for books and ratings of 1 and 5 are dominant in the video games sample.

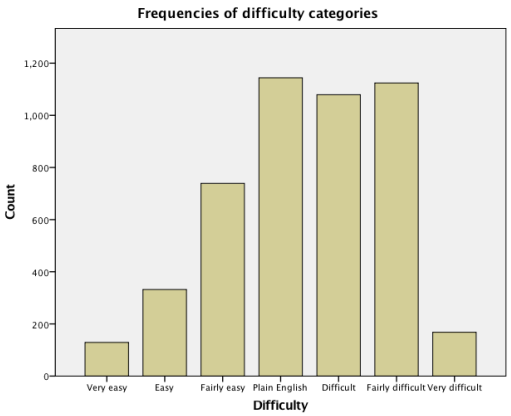

Figure 4.8: Books sample

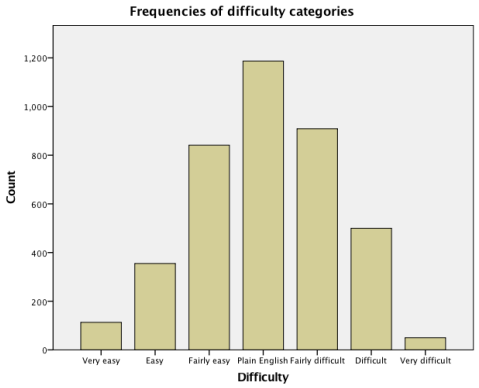

Figure 4.9: Video games sample

Reviews ranging from Plain English to fairly difficult are more numerous for the book sample. Reviews that are fairly easy, plain English, or fairly difficult are most frequent in the video games sample.

4.2.2 Search goods

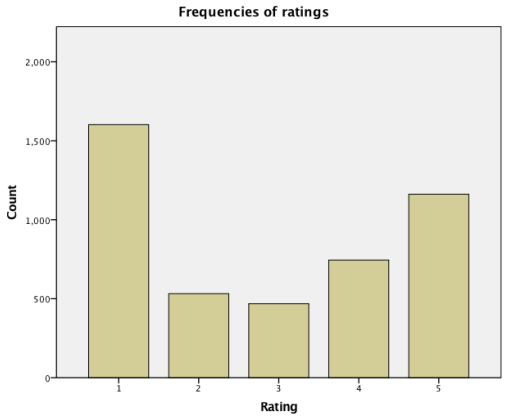

Figure 4.10: Cell phones sample

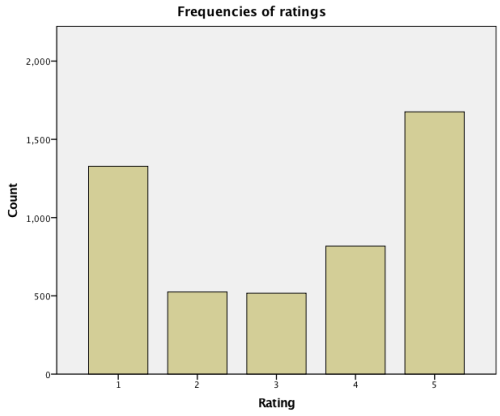

Figure 4.12: Electronics sample

Negative ratings are a lot more frequent with search goods than with experience goods.

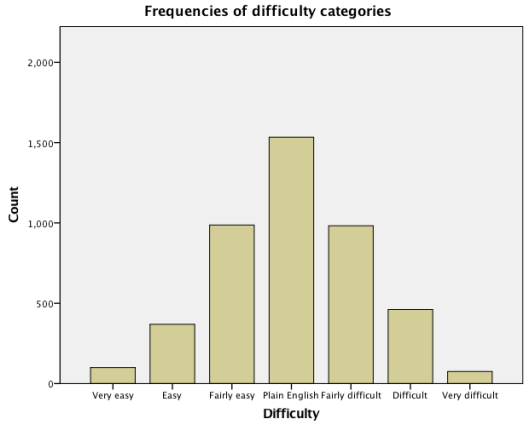

Figure 4.13: Cell phone samples

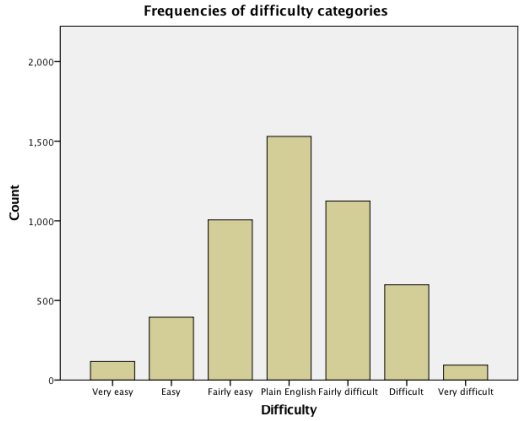

Figure 4.14: Electronics sample

The text difficulty follows the same pattern as experience goods, however plain English texts are distinctively dominant while adjacent difficulties are the second most frequent types of texts.

5 Hypothesis

Based on research and the descriptive statistics we will now make a few hypotheses for the two major product categories.

5.1 Experience goods

- H_1: Reviews with less star ratings are more helpful.
- H_2: Texts with a negative or positive polarity are more helpful.
- H_3: Reviews with more difficult texts are more helpful than easy reviews. However, reviews with extreme text difficulties are less helpful.

5.2 Search goods

- H_1: Reviews with higher star ratings are more helpful.
- H_2: Neutral texts are more helpful than reviews with a positive or a negative polarity.
- H_3: Reviews with a text difficulty of plain English, fairly easy, and fairly difficult are more helpful.

6 Statistical method used

We used a binomial logistic regression to attempt to model the relationship between dependent and independent variables.
The outcome variable, in this case helpfulness, had to be discrete. Reviews that had a 60% or more helpfulness rating were considered helpful while those under 60% unhelpful.
The explanatory variables were:

- Polarity
- Text difficulty
- Rating of product

Logistic regression models are more flexible than simple linear regression and have certain advantages to linear regression.

A logistic regression doesn't have to have normally distributed residuals, doesn't have to be a linear relationship between variables, and homoscedasticity is not needed.

One of the main disadvantages of this statistical method is the need for large amount of data, however this problem is not an issue in the present case.

With a logistic regression the goal is to try and determine in what category a review falls into, helpful or unhelpful based on independent variables.

There are still assumptions that must be tested in order for the model to be valid.

7 Assumptions

7.1 Assumption #1

The dependent variable has to be dichotomous.

Helpfulness is a dichotomous variable. Two outcomes are possible, either a review is helpful or unhelpful. The threshold is 60%. Reviews where 60% or more people found the review helpful are considered helpful.

7.2 Assumption #2

Independent variables have to be nominal or continuous.

Polarity is a continuous variable. Text difficulty and rating are both ordinal variables however they will be treated as nominal (categorical) variables with more than two categories.

7.3 Assumption #3

Sample size should be large. Logistic regression is based on maximum likelihood estimation. The smaller the sample the less accurate the model will be.

This assumption is met. Data size for each sample is bigger than the minimum requirement (50 observations).

	N	Helpful	Unhelpful	Helpful %	Unhelpful %
Books	4715	2480	2235	52.6%	47.4%
Video Games	3955	2166	1789	54.8%	45.2%
Electronics	4864	2376	2488	49%	51%
Cell Phones	4509	2261	2248	50.14%	49.86%

Figure 7.1

All the samples are more or less divided equally between helpful and unhelpful reviews, with search goods being closer to this proportion.

7.4 Assumption #4

There should be independence of observations.
Each review for a same product should be independent from one another. A reviewer should normally be able to make one review per product, he also normally writes his reviews based on his experience of the product and not based on what other reviewers wrote.

We can assume that reviews are independent from one another.

7.5 Assumption #5

There needs to be a linear relationship between independent variables and the logit transformation of the dependent variable. We used a Box-Tidwell approach to test for linearity. It's a two-step approach.

1. The continuous independent variables are transformed into their natural logs.
2. Create interaction between the independent variables and their natural logs.

This assumption doesn't apply to nominal variables so the variables rating and text difficulty won't need to be tested.

Because a natural log transformation is needed we first did a transformation to the polarity variable. The LN transformation doesn't work on negative values and zeros. To avoid this problem, we simply added a constant equal to the minimum value of the sample plus 1.

$$k = |Polarity(\min)| + 1$$

The transformation doesn't modify distribution.

Concerning the p-value, the standard p value is $p<0.05$. Because there are three variables we have to modify the p-value at which statistical significance is accepted by dividing it by the number of independent variables, in this case by 3. This gives us a new p value of 0.01667.

The interaction term should be bigger than 0.01667 in order for there to be linearity between the logit of the dependent variable and the independent variable.

7.5.1 Books

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Transformed Polarity	-.125	.094	1.798	1	.180	.882
	LN(Transformed Polarity) by Transformed Polarity	.023	.019	1.413	1	.234	1.023

Figure 7.2

The assumption is respected. The transformed polarity is linearly related to the logit of the dependent variable.

7.5.2 Video Games

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LN(Transformed Polarity) by Transformed Polarity	.024	.026	.813	1	.367	1.024
	Transformed Polarity	-.119	.136	.776	1	.378	.887

Figure 7.3

Polarity is linearly related to the logit of the dependent variable.

7.5.3 Electronics

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LN (Transformed Polarity) by Transformed Polarity	.140	.029	23.138	1	.000	1.150
	TR_Polarity	-.587	.124	22.234	1	.000	.556

Figure 7.4

The interaction term is statistically significant. The independent variable polarity failed the assumption of linearity.

In order to keep this variable in the equation it will be transformed into a categorical variable.

Category	Polarity score
Positive	1 and higher
Neutral	0
Negative	-1 and lower

7.5.4 Cell Phones

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LN(Transformed Polarity) by Transformed Polarity	.139	.035	15.971	1	.000	1.149
	Transformed Polarity	-.632	.153	16.957	1	.000	.532

Figure 7.5

The cell phone failed the box Tidwell test. This variable will, like in the previous model, be transformed into a categorical variable.

7.6 Assumption #6

There shouldn't be any multicollinearity between independent variables. To test for multicollinearity we examined tolerance and variance inflation factor (VIF). If the tolerance is lower than 0.1 or the VIF higher than 10 there may be multicollinearity.

7.6.1 Books

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Polarity	.892	1.121
	Difficulty	.948	1.055
	Rating	.939	1.064

Figure 7.6

None of the variables have a tolerance under 0,1 and a VIF over 10. There isn't any multicollinearity between the explanatory variables.

7.6.2 Video Games

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Polarity	.913	1.096
	Difficulty	.948	1.055
	Rating	.907	1.103

Figure 7.7

None of the variables have a tolerance under 0,1 and a VIF over 10. There isn't any multicollinearity between the explanatory variables.

7.6.3 Electronics

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Polarity	.853	1.172
	Rating	.847	1.181
	Difficulty	.989	1.011

Figure 7.8

There isn't any multicollinearity between independent variables.

7.6.4 Cell phones

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Polarity	.891	1.122
	Rating	.888	1.126
	Difficulty	.990	1.010

Figure 7.9

There isn't any multicollinearity for the cell phones sample.

7.7 Assumption #7

There shouldn't be any significant outliers that have a high impact on the model.

7.7.1 Books

There were 46 outliers (1%). All the outliers were reviews the model predicted as helpful but were in fact unhelpful.

7.7.2 Video Games

There were a few outliers for this data set, 34 were detected (0.9% of the data).

7.7.3 Electronics

There were 49 outliers (1% of the data). All observations had reviews predicted as unhelpful when in reality they were helpful.

7.7.4 Cell Phones

This is the data set with the least outliers: 27 (0.6%). We observe the same thing as with electronic goods, reviews were predicted unhelpful when the actual observed helpfulness was helpful.

7.8 Origin of outliers

Because outliers were found in all samples it is important to find out why these observations don't fit.

The first thing we did was look at the value of each independent variable for each outlier. We found there was a pattern (see annex for table).

7.8.1 Search goods

The search good outliers acted in the same way for both samples.

The two factors that seem to influence outliers are rating and text difficulty.

If we look at the bar charts below we can see that as ratings go up so does helpfulness. The outliers for cell phones all had 1 or 2 stars and the outliers for electronics had 1,2 or 3 stars. Naturally the model expected these reviews as being unhelpful when they were actually helpful reviews.

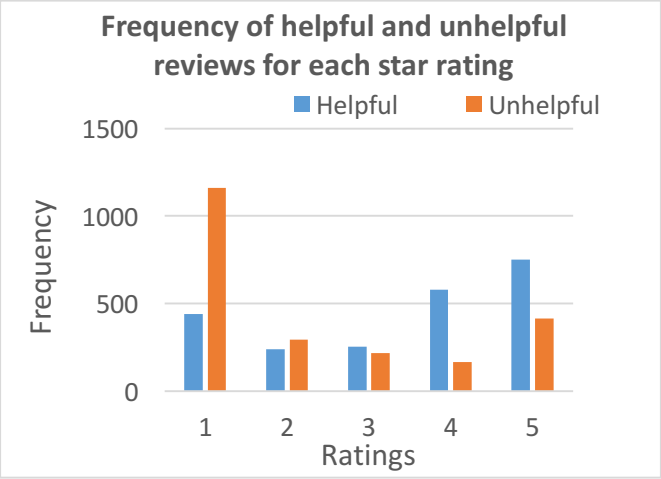

Figure 7.10: Cell Phones sample

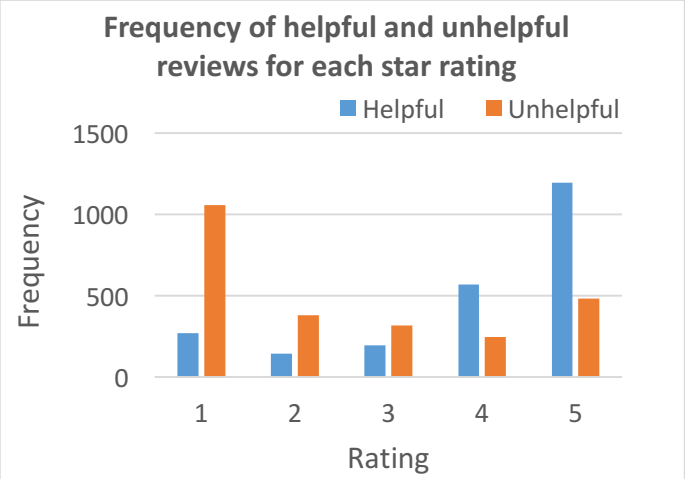

Figure 7.11: Electronics sample

The rating isn't the only influential factor. The text difficulty also played a role.

We can take a look at the graphs below. Reviews with a text difficulty of 1, 5, 6 or 7 (very difficult, fairly easy, easy, and very easy) tend to be less helpful than other reviews. All outliers had either a grade of 1, 5, 6, or 7 for electronics and a grade of 1, 6, or 7 for cell phones.

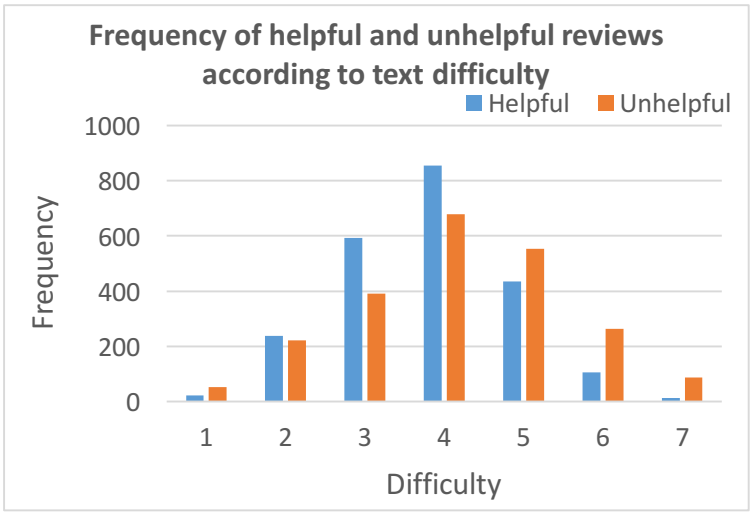

Figure 7.12: Cell Phones sample

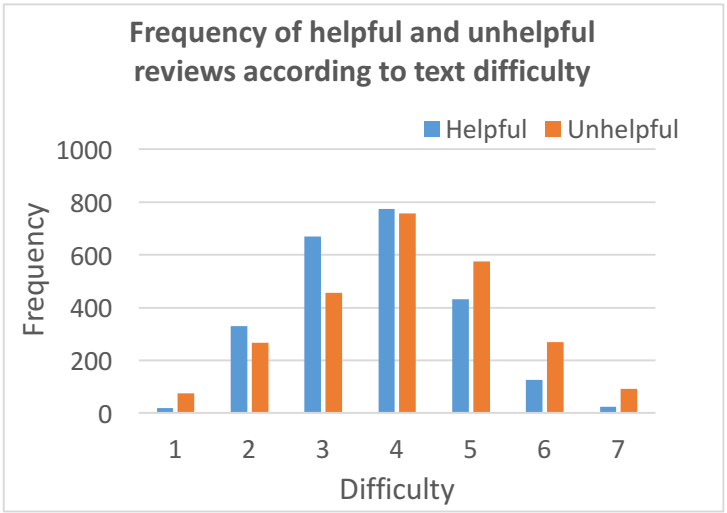

Figure 7.13: Electronics sample

The combination of these two factors is what created outliers in the search good samples. The model expected reviews to be unhelpful when in fact they were helpful.

7.8.2 Experience goods

We will first take a look at the video game sample. The model predicted for these outliers both unhelpful and helpful reviews.

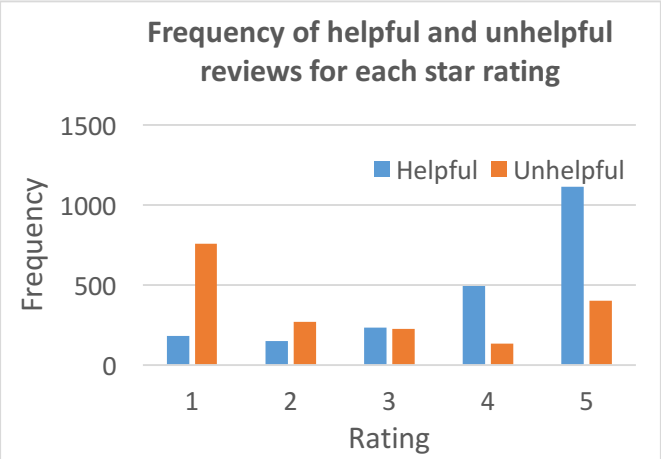

Figure 7.14: Video Games sample

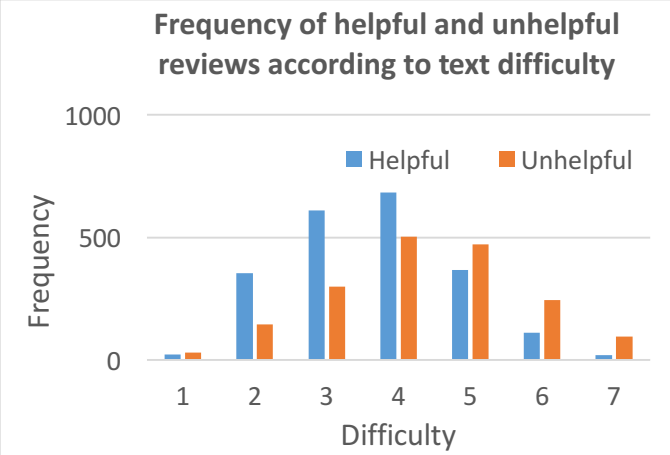

Figure 7.15: Video Games sample

If we look at the two graphics we can see that the tendency is the same as with search goods. As ratings go up so does helpfulness and as texts become more difficult, helpful reviews become more prominent.

Outliers that were predicted as unhelpful had ratings of 1 or 2 and difficulties of 6 or 7. Those who were predicted as helpful had ratings of 4 and text difficulties of 2.

It is possible that "video games" could be classified as a search good rather than an experience good. Or it could simply be that technological goods or goods with "short life spans" have the same tendency.

The last sample, books, was influenced by the same factors. All outliers were predicted as helpful when the actual observations were unhelpful.

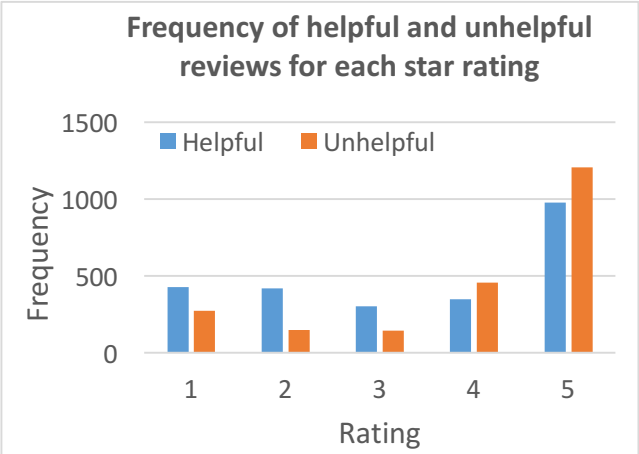

Figure 7.16: Books sample

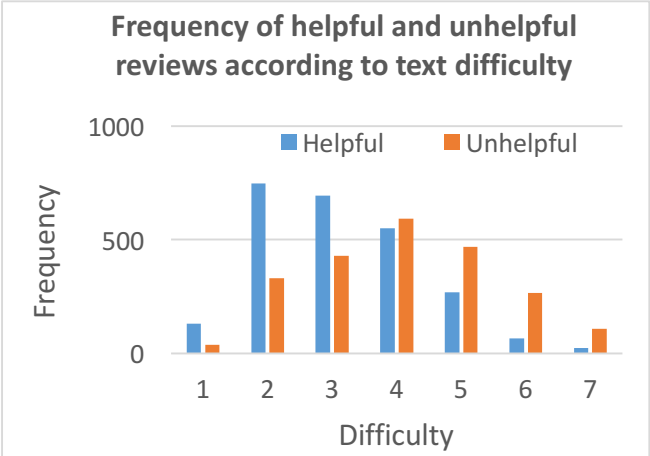

Figure 7.18: Books sample

Figures 5.16 and *5.17* give us the general tendency of the two independent variables compared to helpfulness. We can see that as ratings go down helpfulness goes up. As texts become more difficult reviews become more useful.

All outliers had ratings of 1, 2 or 3 and text difficulties of 1 or 2. The model predicted them as helpful when in fact they were unhelpful.

7.9 Decision concerning assumptions

The outliers will be kept in the final models. It is possible that these observations are outliers because they are influenced by other variables. Adding more variables in the model could make them disappear.

We will only use two samples for the final model: books and cell phone samples. The book sample seem to act differently from search goods while the video games sample appears to act the same way. We chose the cell phone sample because it has less outliers than the electronic sample.

NB: Results for the electronic and video games samples are available in the annexes.

8 The models

8.1 Experience good: Books

8.1.1 Baseline analysis

Classification Table[a,b]

Observed			Predicted: Helpfulness		Percentage Correct
			Unhelpful	Helpful	
Step 0	Helpfulness	Unhelpful	0	2235	.0
		Helpful	0	2480	100.0
	Overall Percentage				52.6

Figure 8.1

This table shows that without any independent variables added the best educated guess is to say that a review is helpful. If you assume this, you have a 52.6% chance of being correct.

8.1.2 Model fit

8.1.2.1 Omnibus tests of model coefficients

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	788.425	11	.000
	Block	788.425	11	.000
	Model	788.425	11	.000

Figure 8.2

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5735.217[a]	.154	.205

Figure 8.3

This table provides an overall look at the statistical significance of the model with independent variables.

The "Model" line is what interests us, it compares the baseline model with the model that has independent variables.

The model is significant, adding the explanatory variables improves the accuracy of the model. The independent variables influence the dependent variable.

The Nagelkerke R Square tells us that the model explains 20.5% of the variation in the outcome.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	152.515	8	.000

Figure 8.4

The Hosmer and Lemeshow test is another way to test the goodness of fit. If the test is statistically significant, as is the case here, then the model is a poor fit.

8.1.3 Model prediction

Classification Table[a]

			Predicted		
			Helpfulness		Percentage Correct
	Observed		Unhelpful	Helpful	
Step 1	Helpfulness	Unhelpful	1328	907	59.4
		Helpful	584	1896	76.5
	Overall Percentage				68.4

a. The cut value is .500

Figure 8.5

This table tells us that with the explanatory variables the model classifies the outcome correctly 68.4% of the time. This is an improvement to the null model (*figure 6.1*) who had a 52.6% rate.

8.1.4 Variables in the equation

NB: each categorical variable has a reference category to which it is compared.

	Reference categories
Difficulty	Very easy
Rating	One star
Polarity	Neutral

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Rating			203.219	4	.000			
	2 stars	.622	.129	23.251	1	.000	1.863	1.447	2.399
	3 stars	.220	.134	2.710	1	.100	1.246	.959	1.619
	4 stars	-.830	.112	55.093	1	.000	.436	.350	.543
	5 stars	-.611	.096	40.894	1	.000	.543	.450	.654
	Difficulty			404.430	6	.000			
	Very difficult	2.859	.305	87.993	1	.000	17.450	9.601	31.715
	Difficult	2.306	.249	86.110	1	.000	10.036	6.166	16.335
	Fairly difficult	1.909	.247	59.731	1	.000	6.749	4.158	10.952
	Plain English	1.346	.246	29.903	1	.000	3.843	2.372	6.227
	Fairly easy	.915	.251	13.305	1	.000	2.496	1.527	4.082
	Easy	.107	.277	.149	1	.699	1.113	.647	1.914
	Polarity	-.015	.004	13.490	1	.000	.986	.978	.993
	Constant	-1.179	.251	22.101	1	.000	.308		

Figure 8.6

This table gives the contribution of each independent variables to the model as well as its statistical significance.

8.1.4.1 Rating

We can see that the rating variable is overall significant (Wald=203.219, df: 4, p<.000).

By looking at the Exp(B) column we can see that reviews with two stars have 1.863 times more chance to be helpful than reviews with one star (reference category). Reviews with five stars are 0.543 times less likely to be helpful than reviews with on star. Reviews with 3 stars aren't significant and reviews with 4 or 5 stars are less likely to be helpful than reviews with 1 star.

8.1.4.2 Difficulty

Overall the difficulty variable is significant (Wald=404.430, df=6, p<.000). The easy category isn't statistically significant. This is the variable that contributes the most to the model.

A very difficult review is 17.450 times more likely to be helpful than a very easy review (reference category).

8.1.4.3 Polarity

Polarity is significant (Wald=13.490, df=1, p<.000), although it contributes the least to the model. An increase in one unit of polarity causes the odds of a review being helpful to decrease by 0.986. Or, in other words, a decrease in one unit of polarity increases the odds of having a helpful review by 1.014 (1/0.986). It seems that the more "negative" a review is the more likely it is to be helpful.

8.2 Summary of experience good models

8.2.1 Books

The logistic model was statistically significant Chi-square=788.425, p<.000. The model explained 20.5% of the variance. With the independent variables added the logistic models precision was of 68.4%. All explanatory variables were statistically significant (figure 6.5). Reviews are less likely to be helpful as text becomes easier. As text difficulty increases helpfulness increases. As ratings go up helpfulness decreases. The more negative a review is the more likely it is to be helpful.

8.2.2 Video games

The model was statistically significant (Chi square=1168.010), p<.000. With the independent variables added there was an increase of 19.6% in precision (54.8% to 74.4%). The model explains 34.2% of the variance.

Out of the three variables only text difficulty and rating were significant. As texts become more difficult helpfulness increases. Reviews who have ratings under 5 stars tend to be unhelpful while reviews with 5 stars are more often helpful than reviews with one star.

8.3 Search goods: Cell Phones

8.3.1 Baseline analysis

Classification Table[a,b]

Observed			Predicted: Helpfulness		Percentage Correct
			Unhelpful	Helpful	
Step 0	Helpfulness	Unhelfpul	0	2247	.0
		Helpful	0	2261	100.0
	Overall Percentage				50.2

Figure 8.7

When dealing with a cell phone product you have a 50/50 chance of the review being helpful or unhelpful if you don't take into account any independent variables.

8.3.2 Model fit

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	895.393	12	.000
	Block	895.393	12	.000
	Model	895.393	12	.000

Figure 8.8

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5353.979[a]	.180	.240

Figure 8.9

The model is statistically significant (p<.0005).

The Nagelkerke R Square tells us that the model explains 24% of the variation in the outcome.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.899	8	.084

Figure 8.10

The result of the test isn't statistically significant indicating a good fit.

8.3.3 Model prediction

Classification Table[a]

			Predicted		
			Helpfulness		Percentage Correct
	Observed		Unhelpful	Helpful	
Step 1	Helpfulness	Unhelpful	1547	700	68.8
		Helpful	666	1595	70.5
	Overall Percentage				69.7

a. The cut value is .500

Figure 8.11

There is an increase of 19.5% in the precision of the model. The model categorizes reviews correctly 69.7% of the time with the explanatory variables added.

8.3.4 Variables in the equation

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Rating			545.131	4	.000			
	2 stars	.761	.106	51.233	1	.000	2.140	1.737	2.635
	3 stars	1.112	.112	99.236	1	.000	3.039	2.442	3.782
	4 stars	2.260	.110	419.995	1	.000	9.580	7.718	11.891
	5 stars	1.677	.091	338.953	1	.000	5.352	4.477	6.398
	Difficulty			159.921	6	.000			
	Very difficult	.843	.411	4.209	1	.040	2.323	1.038	5.197
	Difficult	1.874	.325	33.191	1	.000	6.511	3.442	12.316
	Fairly difficult	2.299	.317	52.466	1	.000	9.967	5.350	18.569
	Plain English	2.159	.314	47.150	1	.000	8.666	4.679	16.051
	Fairly easy	1.676	.317	27.994	1	.000	5.345	2.873	9.945
	Easy	1.059	.333	10.141	1	.001	2.884	1.503	5.535
	Polarity			16.219	2	.000			
	Positive	.314	.122	6.631	1	.010	1.368	1.078	1.737
	Negative	.467	.119	15.328	1	.000	1.596	1.263	2.016
	Constant	-3.275	.328	99.406	1	.000	.038		

Figure 8.12

8.3.4.1 Rating

The rating variable is significant, it increased the accuracy of the model (Wald=545.131, df=4, p<.000). It is the variable that contributes the most to the model. The more stars there are the more likely a review will be helpful. A two-star rating, for example, is 2.14 times more likely to be helpful than a 1-star rating (reference category). The 4-star rating is most likely to be associated with a helpful review. A review with 4 stars has 9.580 times more chance of being helpful than a review with one star.

8.3.4.2 Difficulty

The difficulty variable is significant (Wald=159.921, df=6, p<.000). It is in second place in terms of contribution to the model. We can observe that reviews that are too easy or too difficult tend to be less helpful. A fairly difficult review, best score (Exp(B)=9.967), is 9.967 times more likely to be helpful than a very easy review (reference category).

8.3.4.3 Polarity

Polarity is a significant variable (Wald=16.219, df=2, p<.000), although it contributes the least to the model. A positive review has 1.368 more chance of being helpful than a neutral review and negative review is 1.596 more likely to be helpful than if the review was neutral (reference category).

8.4 Summary of search good models

8.4.1 Cell phones

The model was statistically significant (Chi-square=895.393, p<.000). The model explained 24% of the variance and there had a 69.7% accuracy. All three variables were increased model accuracy. As ratings go up the number of helpful reviews with ratings of 4 being the most likely to have helpful reviews. Fairly easy and plain English texts had the most chance of being helpful. Positive and negative reviews increased the likelihood of a review being helpful.

8.4.2 Electronics

The model was statistically significant (Chi-square=1262.649, <.000). The model explained 30.5% of the variance. There was a 22.2% boost in model prediction accuracy with the explanatory variables added (51.2% to 73.4%). The three independent variables were statistically significant and acted the same way as cell phones sample.

8.5 Hypothesis results

8.5.1 Experience goods

Concerning the book sample, the results seem to confirm the hypothesis.

- Reviews with 1 star are more helpful than reviews with more stars
- Texts with a polarity instead of being neutral seem to be more helpful, although polarity plays a less important role than the two other variables.
- Reviews with a higher difficulty level are more helpful. The model confirms this hypothesis

Unlike the book sample the video games sample doesn't confirm all the hypothesis on experience goods.

- Reviews that had less stars where less helpful than reviews with more stars. This does not support the hypothesis number 1.
- Polarity did not play a significant role in the model
- The harder a text was the more helpful it was; however very difficult reviews helpfulness were less helpful than difficult reviews.

8.5.2 Search goods

Both samples acted the same way. They didn't completely confirm all assumptions:

- Both samples confirmed that reviews with more stars are more helpful.
- They didn't confirm that neutral texts are more helpful. It is however possible that transforming this variable into a nominal variable may have biased results. Although the variable increased model accuracy, it may have been better to not use the variable at all and use the number of words or number of sentences instead. These variables may have yielded better results.
- The test difficulty hypothesis proved partially right, fairly difficult reviews seems to be the most helpful kind of reviews.

9 Conclusion

The different models showed that the explanatory variables had a significant impact on helpfulness.

	Experience goods		**Search goods**	
	Books	**Video games**	**Cell phones**	**Electronics**
Model accuracy without independent variables	52,6%	54,8%	50,2%	51,2%
Model accuracy with independent variables	68,4%	74,4%	69,7%	73,4%
Rating	Reviews with **2 stars** were most likely to be helpful than reviews with 4 or 5 stars.	Reviews with **4 or 5** are more helpful.	Reviews with **4 or 5** are more helpful.	Reviews with **4 or 5** are more helpful.
Text difficulty	The **more difficult** the text the more helpful it is. Extremely difficult texts however tend to be less helpful.	Reviews with **fairly difficult** or **difficult** texts tend to be more helpful. Extreme texts are least helpful.	Helpful reviews usually **have fairly difficult** or **plain English** texts.	Reviews with **fairly difficult** or **difficult** texts tend to be more helpful.
Polarity	The more **negative** a text the more helpful it will be.	The polarity was **not significant**.	Having a **positive** or **negative** review is more helpful than having a neutral review.	Having a **positive** or **negative** review is more helpful than having a neutral review.

Table 9.1: summary of results

Table 9.1 gives an overview of how each explanatory variable influences the helpfulness of a review. First of all, as we can see there is an **increase in the model prediction** with independent variables added. There is on average a **19.28% increase** in the models accuracy when explanatory variables are added.

Concerning the rating, this is the variable that contributes the most for all models except for the book sample, where it is in second place. If we look at the table above, we can see that

the book model is also the only one that doesn't follow the same pattern. Reviews with **4** or **5 stars** are usually helpful for cell phones, video games, and electronics products. Reviews for books on the other hand have to have ratings of **2** to be helpful and reviews that have 4 or 5 stars tend to be unhelpful.

The next variable is **text difficulty**, this is the **second biggest contributor** to the model, except for books where review difficulty is the most important attribute for a review. Once again the book sample is the exception to the rule. For books the **harder the review the more helpful it is.** However, extremely difficult texts are usually unhelpful. For video games and electronics **fairly difficult and difficult reviews are associated with helpful reviews.**

The last variable is **polarity**; it is, for all four models, the **smallest contributor**. In fact, for the video game sample it isn't significant. For the book sample negative reviews are more helpful this goes hand in hand with the rating (lower ratings are more helpful). For the cell phones and electronics models having a positive or negative review will increase the odds of having a helpful review.

Identification of shortcomings in the model is vital in order to increase the model accuracy and usefulness. Here are a few possible improvements.

Firstly, categorizing products into two broad categories may be too simplistic. People who buy books aren't influenced by the same variables as people who buy video games. Analysing what kind of person buys a certain product may help **identify more pertinent criteria** or complementary criteria to the ones already present in the model.

Improving the sentient variable (polarity) may yield better results. Using the lexicon gives a general idea on the polarity of text but isn't extremely accurate. Some reviewers write reviews where they assess product quality as good but say that Amazon service is bad. This can lead to negative polarity when in fact it should be positive. The goal of a product review is to evaluate product quality and not the service provided by Amazon (unless Amazon produces the product).

For search goods it may be interesting to do like Soo-Min Kim, Patrick Pantel, Tim Chklovski, & Marco Pennacchiotti (2006) did and use **product features**. Using product features to identify a positive review from a negative review.

Analysing reviews where a lot of people voted helpful may give better results on what variables effect the helpfulness rating. Reviews who only have 6 or 8 votes may have been voted by marginal consumers who don't have the same criteria as the majority of consumers.

Adding a **control variable** like Susan M. Mudambi and David Schuff (2010) did for the number of votes on a review could improve the model. A review with one vote shouldn't have the same weight as a review with a 100 votes.

Finally, it is important to note the practical application to findings in this domain. Soo-Min Kim, Patrick Pantel, Tim Chklovski, & Marco Pennacchiotti (2006) suggest a **recommendation system**, for example by identifying reviews that a particular user would find useful. They also suggest just simply ranking reviews based on the helpfulness of the text.

Georg Lackermair, Daniel Kailer, & Kenan Kanmaz (2013) take a different point of view by suggesting that it could also help reviewers. By identifying what makes a good helpful review, reviewers can **increase the quality of their review** and create reviews that support the consumer in making their decision.

We suggest a different point of view: the **manufacturers**. Reviews are a gold mine concerning user feed-back. Manufacturers can easily extract the data and find what is wrong with their product, they can also check if the product sells to the target audience or to a completely different audience. By identifying helpful reviews from unhelpful reviews it makes it easier for manufacturers to identify which reviews to take into account and which can be left behind. The number of reviews being humongous this would considerably reduce their amount of work.

10 Annexes

10.1 Syllable code

```
function syllables(text){
    text = text.toLowerCase();
    /* This line transforms the whole review to lowercase so as to avoid miscounts
    due to capital letters */

    text = text.replace(/(?:[^laeiouy]es|ed|[^laeiouy]e)$/, '');
    /*This line checks for silent vowels and removes them. Words that end with
    a consonant (except for l) + "es", the vowel is than removed.*/

    word = text.replace(/^y/, '');
    /*This line removes "y" form begining of words*/

    return text.match(/[aeiouy]{1,2}/g).length;
    /* This line counts the occurences of single and double vowels (diphtongs).
    It then returns the number of syllables*/
}
```

10.2 Descriptive statistics

	Books		**Video games**	
	Words	**Sentences**	**Words**	**Sentences**
N	4715.00	4715.00	3955	3955
Mean	180.12	11.40	170.53	12.45
Std. Error of Mean	2.94	0.17	2.932	0.208
Median	116.00	8.00	112	9
Mode	20.00	3.00	44	4
Std. Deviation	202.02	11.94	184.38	13.106
Variance	40810.38	142.64	33995.818	171.762
Skewness	3.01	3.36	3.759	4.136
Kurtosis	14.63	19.30	29.369	31.745
Range	2280.00	149.00	3148	196
Minimum	7.00	1.00	3	1
Maximum	2287	150	3151	197
	Cell phones		**Electronics**	
N	4508	4508	4864	4864
Mean	157.53	12.12	131.68	9.72
Std. Error of Mean	2.633	0.203	1.935	0.135
Median	103	8	93	7
Mode	20	3	28a	4
Std. Deviation	176.803	13.598	134.923	9.39
Variance	31259.386	184.917	18204.325	88.176
Skewness	4.362	4.77	3.956	3.903
Kurtosis	42.963	51.188	33.572	32.756
Range	3284	281	2475	174
Minimum	5	1	6	1
Maximum	3289	282	2481	175

10.3 Reviews

10.3.1 Review 1

The game is fun to play and the graphic change is tremendous improvement over the original Alpha Centauri;However the time it took to change a few graphics and one or two game points is so little; I was hardly worth buying originally; this is why so few people bought it and why its value is higher;I like the improvements to the interface and story line the alien crossfire yields; but it is weak as far as an improvement; however I'll probably keep my copy. The two alien race factions Manifold Caretakers and the Manifold Usurpers and the terraforming icons for these factions is pure genius and artistically great; my favorite faction is the Nautalis Pirates;And the fungal tower is a refreshing improvement to the creatures of Alpha Centauri;

10.3.2 Review 2

This review had a test score of 60.43, category: plain English

The most wonderful thing about this book is the writing style. The book moved along at a strident pace but it did not lack any detail. It was a great story of assimilating into a different country and culture with the yoke of a name (shared by no one) constantly harnessed about himself. The story was lovely and the writing grand. The characters were so well-defined that I felt like I knew them, like I could easily picture them in my mind. I loved the book and look forward to reading her first book.

10.3.3 Review 3

Test score of 75.99, category: fairly easy

They need to make a better game than this. Only 8 boarders to choose from. Only 7 courses. The one who came up with this must be FIRED!

10.3.4 Review 4

Test score 39.80, category: difficult

The camera takes excellent pictures in just about all conditions, and the autofocus is very reliable; it rarely gets fooled by complicated situations. Flash pictures turn out quite well, with good even lighting throughout the picture. Reliability has been a question, as the camera sometimes gets fussy and won't function correctly. Removing the battery and cleaning the contacts usually fixes the problem, but I have not had to do that before with a camera, and it is too frequent an occurrence with this one. In these cases of malfunction, the battery tests out fine so it is not the problem. If reliability were better this would be a five star camera.

10.3.5 Review 5

Test score of 80.91, category: easy

This item came with poor instructions on how to use it: for the first few days, I couldn't get the cover off. I was afraid to force it, that I might break it. I tried to get help from their web site but was ignored. I guess they thought the question was below them. Fortunately, a friend of mine who is stronger, was able to take the cover off and I could use it.

10.4 Outlier table

10.4.1 Books

Outlier row number	Difficulty	Rating	Polarity	Predicted outcome
676	2	2	-17	H
2337	2	2	-9	H
2566	1	3	-5	H
2885	2	2	-20	H
2903	2	2	-24	H
3252	2	2	-9	H
3400	2	2	-12	H
3489	1	3	-17	H
4037	1	1	-15	H
4076	2	3	-39	H
4337	2	2	-7	H
4350	2	2	-13	H
2478	1	2	-27	H
2485	1	3	2	H
2489	1	2	-24	H
2825	2	2	-18	H
2931	2	2	-13	H
2955	2	2	-12	H
2984	1	3	-2	H
3079	1	3	-7	H
3099	2	2	-7	H
3246	2	2	-33	H
3254	1	2	-1	H
3331	2	2	-9	H
3406	2	2	-19	H
3408	2	2	-12	H
3457	2	2	-8	H
3694	2	2	-20	H
3716	2	2	-10	H
3870	2	2	-9	H
4046	1	1	-17	H
4054	1	1	-35	H

4061	1	1	-13	H
4270	1	2	0	H
4357	2	2	-21	H
4365	1	2	-7	H
4391	2	2	-15	H
4403	2	2	-12	H
4410	2	2	-17	H
4441	2	2	-18	H
4470	2	2	-34	H
4497	2	2	-51	H
4501	2	2	-12	H
4541	1	1	-13	H
4571	2	2	-24	H

10.4.2 Video Games

Outlier row number	Difficulty	Rating	Polarity	Predicted outcome
15	2	4	0	H
713	2	4	5	H
1007	2	4	-8	H
1086	2	4	-6	H
1184	2	4	0	H
1947	2	4	1	H
2297	2	4	-10	H
2629	2	4	1	H
3096	2	4	-9	H
3586	2	4	2	H
3771	2	4	-9	H
3840	2	4	-13	H
815	6	1	-5	U
933	6	1	-6	U
967	6	1	-3	U
1029	6	1	3	U
1218	7	1	-2	U
1227	6	1	-1	U
1397	7	1	2	U
1400	6	1	2	U

1527	7	1	2	U
1727	6	1	-5	U
2169	6	1	-1	U
2215	6	1	1	U
2460	6	1	-4	U
2518	6	1	1	U
2571	7	1	-3	U
2689	6	1	-10	U
2947	6	1	0	U
3077	7	2	-8	U
3193	6	1	0	U
3239	6	1	-8	U
3263	6	1	3	U
3744	6	1	-4	U

10.4.3 Electronics

Outlier row number	Difficulty	Rating	Polarity	Predicted outcome
7	7	1	Positive	U
128	6	1	Negative	U
146	6	2	Neutral	U
177	6	1	Negative	U
182	6	1	Positive	U
593	6	1	Negative	U
754	1	2	Positive	U
824	6	1	Negative	U
1007	5	1	Neutral	U
1119	6	1	Negative	U
1342	6	1	Negative	U
1421	7	1	Negative	U
132	6	1	Positive	U
133	5	1	Neutral	U
139	6	1	Positive	U
156	6	1	Negative	U
193	1	2	Positive	U
266	5	1	Neutral	U

316	7	3	Negative	U
507	6	1	Positive	U
523	1	1	Negative	U
570	6	1	Negative	U
651	1	2	Negative	U
674	1	1	Negative	U
757	6	1	Negative	U
761	6	1	Negative	U
801	6	1	Neutral	U
871	7	1	Negative	U
901	6	1	Positive	U
940	6	1	Positive	U
1019	5	1	Neutral	U
1036	6	1	Positive	U
1042	6	1	Positive	U
1296	7	3	Neutral	U
1483	1	1	Negative	U
1524	6	1	Negative	U
1598	5	1	Neutral	U
1643	6	1	Negative	U
1693	6	1	Negative	U
1788	5	1	Neutral	U
1865	1	1	Negative	U
1884	6	1	Negative	U
2011	7	1	Positive	U
2029	6	1	Negative	U
2032	1	2	Negative	U
2237	5	1	Neutral	U
2267	6	1	Negative	U
2273	1	2	Negative	U
2362	6	1	Positive	U

10.4.4 Cell phones

Outlier row number	Difficulty	Rating	Polarity	Predicted outcome
372	1	1	Positive	U
459	7	2	Negative	U
609	6	1	Positive	U
610	6	1	Neutral	U
1021	6	1	Positive	U
1075	1	1	Neutral	U
1144	6	1	Positive	U
1148	7	1	Positive	U
1186	7	1	Neutral	U
1615	7	1	Positive	U
1706	6	1	Positive	U
1710	6	1	Positive	U
1884	7	1	Positive	U
1889	7	1	Positive	U
1890	6	1	Positive	U
2038	6	1	Neutral	U
2106	1	1	Negative	U
2113	1	1	Neutral	U
2119	7	1	Negative	U
2504	6	1	Positive	U
2540	7	1	Positive	U
2556	6	1	Positive	U
2569	6	1	Neutral	U
3010	6	1	Neutral	U
3048	1	1	Positive	U
3097	6	1	Positive	U
3108	6	1	Positive	U

10.5 Models

10.5.1 Video games

Classification Table[a,b]

			Predicted		
			Helpfulness		Percentage Correct
	Observed		Unhelpful	Helpful	
Step 0	Helpfulness	Unhelpful	0	1789	.0
		Helpful	0	2166	100.0
	Overall Percentage				54.8

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1168.010	11	.000
	Block	1168.010	11	.000
	Model	1168.010	11	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	4278.793[a]	.256	.342

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.821	8	.451

Classification Table[a]

			Predicted		
			Helpfulness		Percentage Correct
	Observed		Unehelpful	Helpful	
Step 1	Helpfulness	Unhelpful	1214	575	67.9
		Helpful	437	1729	79.8
	Overall Percentage				74.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Rating			667.037	4	.000			
	2 stars	.696	.135	26.779	1	.000	2.006	1.541	2.611
	3 stars	1.357	.128	111.657	1	.000	3.885	3.021	4.998
	4 stars	2.648	.133	397.581	1	.000	14.131	10.892	18.333
	5 stars	2.395	.107	501.041	1	.000	10.967	8.892	13.525
	Difficulty			203.166	6	.000			
	Very difficult	.941	.412	5.200	1	.023	2.562	1.141	5.749
	Difficult	2.417	.294	67.578	1	.000	11.215	6.303	19.958
	Fairly difficult	2.294	.284	65.328	1	.000	9.911	5.683	17.286
	Plain English	1.946	.280	48.261	1	.000	7.000	4.043	12.120
	Fairly easy	1.455	.283	26.422	1	.000	4.283	2.460	7.459
	Easy	.767	.300	6.538	1	.011	2.152	1.196	3.873
	Polarity	.003	.005	.234	1	.628	1.003	.992	1.013
	Constant	-3.362	.474	50.270	1	.000	.035		

10.5.2 Electronics

Classification Table[a,b]

			Predicted		
			Helpfulness		Percentage Correct
	Observed		Unhelpful	Helpful	
Step 0	Helpfulness	Unhelpful	2488	0	100.0
		Helpful	2376	0	.0
	Overall Percentage				51.2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1262.649	12	.000
	Block	1262.649	12	.000
	Model	1262.649	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5477.708[a]	.229	.305

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.200	8	.105

Classification Table[a]

Observed			Predicted Helpfulness		Percentage Correct
			Unhelpful	Helpful	
Step 1	Helpfulness	Unhelpful	1824	664	73.3
		Helpful	632	1744	73.4
	Overall Percentage				73.4

a. The cut value is .500

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Rating			781.987	4	.000			
	2 stars	.359	.121	8.827	1	.003	1.432	1.130	1.814
	3 stars	.872	.116	56.153	1	.000	2.392	1.904	3.006
	4 stars	2.183	.107	416.239	1	.000	8.877	7.197	10.949
	5 stars	2.281	.095	581.847	1	.000	9.790	8.133	11.783
	Polarity			16.685	2	.000			
	Positive	.449	.113	15.779	1	.000	1.566	1.255	1.955
	Negative	.425	.115	13.690	1	.000	1.529	1.221	1.915
	Difficulty			152.750	6	.000			
	Very difficult	-.057	.363	.025	1	.875	.945	.464	1.923
	Difficult	1.625	.263	38.226	1	.000	5.077	3.033	8.498
	Fairly difficult	1.926	.256	56.700	1	.000	6.862	4.157	11.329
	Plain English	1.589	.253	39.479	1	.000	4.901	2.985	8.046
	Fairly easy	1.306	.256	25.954	1	.000	3.692	2.234	6.102
	Easy	.799	.274	8.528	1	.003	2.223	1.300	3.800
	Constant	-3.238	.270	143.402	1	.000	.039		

11 References

Articles

Susan M. Mudambi & David Schuff. (2010). What makes a helpful online review ?. *MIS Quarterly, 34,* 185-200

Kumar and Benbasat. (2006). The influence of online product recommendations on consumers' online choices. *Journal of retailing,* 159-169

Michael P. O'Mahony & Barry Smith. (2010). The readability of helpful product reviews. *Florida Artificial Intelligence Society Conference,* 154-155

Soo-Min Kim, Patrick Pantel, Tim Chklovski, & Marco Pennacchiotti. (2006). Automatically assessing review helpfulness. *Empirical methods in Natural Language,* 423-430

Scott Bolter. (2013). *Predicting product review helpfulness using machine learning and specialized classification models* (Master's Theses and graduate research). San José State University, California, USA.

Michael P.O'Mahony & Barry Smyth (2010). The readability of helpful reviews. *FLAIRS,* 154-155

Nikolaos Korfiatis, Elena Garcia-Baricocanal, & Salvador Sanchez-Alonso. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic commerce research and application,* 205-217

Monic Sun. (2011). How does the variance of product ratings matter? 1-28

Philip Fei Wu, Hans van der Heidjen,& Nikolaos Th. Korfiatis. (2011). The influence of negativity and review quality on the helpfulness of online reviews. *International conference on information systems,* 1-10

Nelson, P. (1970). Information and Consumer Behavior. *Journal of Political Economy, 78(2),* 311-329. Retrieved from http://www.jstor.org/stable/1830691

Georg Lackermair, Daniel Kailer, & Kenan Kanmaz. (2013). Importance of online product reviews from a consumer's perspective. *Horizon Research Publishing,* 1-5

Geaorge J. Stigler. (1961). The economies of information. *The journal of political economy, 69,* 213-225

Books

David W. Hosmer Jr., Stanley lemeshow, & Rodney X. Sturdivant. (2013). *Applied logistic regression Third edition*. New-Jersey, USA, John Wiley & sons, Inc.

Websites

Dan Hinckley. (2015). New study: data reveals 67% of consumers are influenced by online reviews. MOZ. Retrieved from https://moz.com/blog/new-data-reveals-67-of-consumers-are-influenced-by-online-reviews

Graham Charlton. (2015). Ecommerce reviews: why you need them and how to use them. Retrieved from https://econsultancy.com/blog/9366-ecommerce-consumer-reviews-why-you-need-them-and-how-to-use-them/

Julian McAuley, Christopher Targett, Qinfeng (Javen) Shi, & Anton van den Hangel. (2015). Amazon product data. Retrieved from http://jmcauley.ucsd.edu/data/amazon/

Phonics on the web. Retrieved from http://www.phonicsontheweb.com/syllables.php

Artfulhacker. (2011). How to compute the number of syllables in a word in javascript? Stackeorflow. Retrieved from http://stackoverflow.com/questions/5686483/how-to-compute-number-of-syllables-in-a-word-in-javascript

Max Woolf. (2014). A statistical analysis of 1.2 million amazon reviews. Retrieved from http://minimaxir.com/2014/06/reviewing-reviews/

Bing Liu. Sentiment symposium tutorial: lexicons. Retrieved from http://sentiment.christopherpotts.net/lexicons.html#opinionlexicon

Adam Lund & Mark Lund. Laerd statistics. Retrieved from https://statistics.laerd.com/aboutus.php

12 Bibliography

Dylan Shinzaki, Kate Stuckman, & Robert Yates. (2013). *Trust and helpfulness in Amazon reviews: Final report*. 1-15

Garret P.Sonnier, Leigh McAlister, & Oliver J. Rutz. (2011). A dynamic model of the effect of online communications on firm sales. *Marketing Science, 30,* 702-716

Wendy W. Moe & Michael Trusov. (2010). *Measuring the value of social dynamics in online product rating forums*. 1-43

Wenjing Duan, Bin Gu, & Andrex B. Whinston. (2008). The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. *Journal of retailing, 84,* 233-242

Chrysanthos Dellarocas, Xiaoquan Zhang, & Neveen F. Awad. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Wiley Periodicals, Inc. and Marketing Educational Foundation, Inc. 21,* 23-45

Yumbo Cao. Assessing quality of product reviews. 1-2

Julian McAuley, Christopher Targett, Qinfeng (Javen) Shi, & Anton van den Hangel. (2015). *Image based recommandations on styles and substitutes*. UCSD, USA, 1-10

Anindya Ghose and Panagiotis G. Ipeirotis. (2009). Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*. 1-15

Executive summary

The purpose of this thesis is to try and create a model capable of **predicting if a review will be helpful or unhelpful** based on explanatory variables.

Four samples were used and were divided into two major groups: **experience goods and search goods.**

The following table summarizes our results.

Independent variable	Experience goods		Search goods	
	Books	**Video games**	**Cell phones**	**Electronics**
Rating	Reviews with **2 stars** were most likely to be helpful than reviews with 4 or 5 stars.	Reviews with **4 or 5** are more helpful.	Reviews with **4 or 5** are more helpful.	Reviews with **4 or 5** are more helpful.
Text difficulty	The **more difficult** the text the more helpful it is. Extremely difficult texts however tend to be less helpful.	Reviews with **fairly difficult** or **difficult** texts tend to be more helpful. Extreme texts are least helpful.	Helpful reviews usually **have fairly difficult** or **plain English** texts.	Reviews with **fairly difficult** or **difficult** texts tend to be more helpful.
Polarity	The more **negative** a text the more helpful it will be.	The polarity was **not significant**.	Having a **positive** or **negative** review is more helpful than having a neutral review.	Having a **positive** or **negative** review is more helpful than having a neutral review.

The following explanatory variables **contributed the most** in our model: rating and text difficulty. Polarity on the other hand did not increase the accuracy of the model significantly.

The models can **contribute to e-commerce** websites where product reviews are frequent. Identifying helpful reviews can help customers make their purchase choice.